SP-2714

AD 648755

SP *a professional paper*

Notes on Semantic Discourse Structure

Karen Sparck Jones

3 March 1967

SYSTEM
DEVELOPMENT
CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

SDC

FOREWORD

This paper is intended to be part of a larger work on the nature of semantic classifications and their role in discourse analysis; it should therefore be regarded as a first version of my views on the topics with which it deals, rather than as a final and fully worked-out one.

ABSTRACT

The semantic problems of natural language discourse analysis are more serious than the syntactic ones, and much less work has been done on them. It is widely held that a semantic classification of words by their meanings is required for selecting the correct senses of individual words in text, but the implications of this hypothesis have not been fully worked out. It is moreover evident that semantic discourse analysis does not end with disambiguation: understanding a piece of discourse depends on identifying its message, and it is arguable that for this we have to know what the semantic structure of the text as a whole is, where this may not be very directly related to the syntactic structure of its component sentences. It is suggested that a semantic or conceptual classification plays a part here, since understanding the message of a text involves some knowledge of the way concepts may be or are usually combined. In this paper, different aspects of this notion of semantic discourse structure are explored, and their connection with a semantic classification is exhibited.

NOTES ON SEMANTIC DISCOURSE STRUCTURE

Introduction: The Use of a Semantic Classification for Disambiguation in Discourse Analysis

We start with the assumption that a semantic classification of the words in a vocabulary by their meanings is required for the semantic analysis of natural language text. The argument for this assertion is basically that any attempt to select the correct senses of words in text by looking for other specific words is so inadequate, both practically and linguistically, that it cannot be the normal means of resolving ambiguity. It is much more reasonable to suppose that we rely on the conceptual character of the context surrounding a word, and that we select the correct sense of the word because it conveys a concept which fits in with those associated with the surrounding context; where we say that it fits in if the resulting conceptual combination is a standard or permitted one, in the sense that it underlies many individual texts containing particular words and so indicates the kind of thing that can be said. We therefore require dictionary entries for the words in our vocabulary noting the general concepts which they convey, in terms of semantic markers or semantic class headings, such as MOTION for "walk," "run," and "jump." And we select the correct sense of a word by looking at the class entries for others in its context to see whether any of the alternative class headings defining their various senses will give us a permitted combination. Thus, to give a highly schematic example, for the sentence "My aunt is chewing rock,"* we select the correct sense of "rock," which is specified by FOOD, as opposed to the wrong sense specified by EARTH, because one heading for "chewing" is EAT, and EAT and FOOD as opposed to EAT and EARTH is a permitted (and well known) combination of ideas. This selection process is thus in principle a mutual one: the correct senses of several words are picked out by considering each of their dictionary entries in relation to the others in the search for a permitted combination.

The foregoing argument can be spelled out at length, but I shall not do this here: it is sufficient to note that it underlies a variety of approaches to discourse analysis, like those discussed by Katz/Fodor/Postal, Weinreich, or some members of the Cambridge Language Research Unit. What I want to consider are some implications of this very general suggestion that a semantic classification is required for discourse analysis. Thus it is obvious, for example, that if our permitted combinations are expressed in terms of conceptual headings, or semantic markers, which we use to classify our words, how we classify our words, and how we represent our combinations are interrelated problems. We might say, for instance (to give a very crude illustration),

*In English English as opposed to American English, "rock" can mean either stone or candy. This ambiguity is essential to the whole argument: I am assuming for the purpose of the example that the correct meaning of "rock" in the sentence is candy.

that "running" suggests the idea of MOTION, or that of SPEED, and that "boy" suggests MALE or PERSON, so that we could look at a sentence like "The boy is running" as having to do either with PERSON and SPEED, or with MALE and MOTION, and so on: and how we classify our words and indicate our permitted combination with these alternatives may make all the difference between successful ana unsuccessful selections of the sense of the words in the sentence. For if we have PERSON and MALE as classes, but list only MALE and SPEED as a permitted combination, then classifying "boy" under PERSON alone would mean they we could not analyze our sentence correctly.

The relation between our class description for words and our set of permitted combinations is, however, more complicated than this, since whether or not we have a permitted combination of ideas may depend on how they are combined; in other words, we compare the class dictionary entries for two words to see if they may go together if they are interpreted in one way rather than the other, and there may be constraints on which words may be compared. The simplest case is where we imagine that a permitted combination is a simple concatenation of headings, and we can look anywhere in a text to find it. But this is certainly too crude, and as soon as we consider something more restricted, we have to accept that what combination we look for may be influenced by where we look--that is, that the combinations we have in our list of permitted combinations will be influenced by what we regard as legitimate comparisons between words. Thus, we may use different information if we attempt to select the correct sense of a word by looking at the entries for any other words in a whole sentence, say, or by looking only at proximate words, or ones which are closely related syntactically to the given word.

How, then, are we to proceed in the investigation of our two problems of specifying classes and combinations, given the way in which they interlock with one another? In general, some division of interest is characteristic of work in this field: thus on the one hand, we find studies of the way in which semantic headings, or markers, may be used in determining sentence meaning, where the emphasis is on the procedure for text analysis and there is little or no attempt to go into the problems of selecting and assigning the markers required to construct the dictionary entries for the items in the vocabulary. And on the other, we find investigations of the relations between the items in a vocabulary, and the classes into which they fall, or the headings under which they may be subsumed, without very much reference to the part which this classification may play in discourse analysis. In some cases, this may simply be the consequence of a restricted aim: thus we may only be concerned with the nature of vocabulary structure and, say, with describing semantic fields, without having any further purpose in mind. In other cases, some more or less plausible assumptions are made about those aspects of the overall problem in which we are not primarily interested. Thus writers on discourse analysis tend to assume that the semantic classification of any word, given that the need for it is recognized, will be as straightforward as that exemplified by the use of headings like HUMAN and

MALE to describe "boy"; while writers who concentrate on the question of classification tend to behave as if the resulting descriptions of words will fit, for example, into schedules of permitted combinations of the kind mentioned earlier, in the obvious way in which EATING and FOOD for "chewing" and "rock" form a combination.

Either of these alternatives is basically unsatisfactory. A more important point, however, is that we can only concentrate on one problem or the other when we have a fairly good idea of the role of a semantic classification in discourse analysis. The argument with which we began gives one reason for using a semantic classification, in connection with one part of discourse analysis, namely, disambiguation; but discourse analysis is not simply a matter of disambiguation, and there may be other uses for a semantic classification. In what follows, therefore, I shall consider some semantic features of discourse, and whether they are associated with semantic classification, in more detail.

Other Semantic Features of Discourse and Other Uses of a Classification

If we look at the work now being done which is primarily concerned with the semantic aspects of discourse, we find that this ranges from purely descriptive discussions of noticeable features of texts, to attempts to formulate more or less strong models of the nature of discourse and to apply them to the actual analysis of text. Such models, or at least sets of rules, may also be tested by being used not to analyze given text, but to construct it. In the first instance, such approaches tend to be associated with attempts to identify the semantic character of a sentence, and this is especially the case where the semantic procedure is closely associated with, or heavily dependent on, the detailed syntactic description of pieces of text, since this emphasizes sentence boundaries. But much of the work on this subject is concerned with the semantic properties of larger pieces of text, up to, say, paragraph limits.

Extra-Sentential References and Disambiguation

The reason for this interest in larger pieces of text is obvious. In the discussion of "My aunt is chewing rock" I behaved as if there was only one genuine semantic interpretation of the sentence, namely that my aunt was eating candy. But she could have been struggling with a piece of stone; and the fact that this is improbable, in that we assume that the first interpretation is the natural one, does not matter here, since we are interested only in whether alternative semantic interpretations are possible, though they may be more or less unlikely: so long as alternative interpretations are possible in principle, we have to try to find some means of selecting one of them. It may be that we assume that the words in a sentence do have one meaning rather than another, in the way in which we would probably say that "rock" in our sentence means candy and not stone: so that we might say that we want to confirm this selection, rather than that we want to make this selection. On this view, that is, we are concerned with obtaining sufficient information to confirm a tentative selection, and not to make a selection. But whether we look at the situation in one way or the other, we are basically dealing with only one problem, namely, that of searching for further information so that one sense of a word rather than another may be identified as the correct one. For convenience, therefore, I shall treat the problem as one of resolving ambiguity, of making a selection, rather than of confirming some choice, but nothing that I shall say should preclude the other view, and the latter may indeed be more natural in some contexts.

Given, therefore, that many sentences are semantically ambiguous in principle because at least one word in them may be interpreted in more than one way, how is this ambiguity to be resolved? It can only be done by using more information from the surrounding linguistic context. We started by considering the immediate surroundings of a word within a sentence, and we now have to consider the larger textual context surrounding the sentence. It seems clear, moreover, that the additional information which may be obtained from looking at this larger context should be of the same kind as that derived from the sentence alone, and that it should be used in the same way. Thus, for example, if we were working with a very simple repetition model in which the correct senses of words are selected by recurring concept headings, we would look for recurring headings in the surrounding text; and if we were looking for permitted combinations of a less restricted kind, we would look for the components of such combinations in the larger text.

Now the assumption behind the foregoing is that if we look at an extended piece of discourse, we will be able to obtain the necessary information. Consider, for instance the following text:

"Mr. Smith is a big man. He went to the bank. He did his best work there. His object was good. It had a point."

The individual sentences are semantically ambiguous, and no means of resolving these ambiguities are supplied. For this reason alone, this text looks queer; and it is an empirical fact that texts are not normally like this. That our assumption is plausible, and that we can find the kind of information we are looking for in the larger context is shown, for example, by an ordinary piece of text, like the following paragraph from an economic history textbook:

"About the middle of the eighteenth century, the communications of Great Britain lagged behind general economic development. The only age of systematic road building which the island had ever known had been under the Romans. The canal and railway eras were as yet in the future. Since the close of the Civil War, some improvements in communications both by road and water had been carried out. On the roads the first turnpike trusts, instituting a new system by which local companies raised loans for repair and maintenance, which they paid back out of a revenue derived from tolls upon the traffic, dated from the early eighteenth century. The turnpikes represented a method of financing and managing the roads which might succeed where the parish had failed."

Here, we find, throughout the text as a whole, both repetitions of some headings like ROUTE, and also occurrences of related headings like COMMUNICATION, TRAVEL and TRANSPORT. The fact that we may find the necessary information somewhere in the surrounding text does not, however, help us very much when we try to set up rules for searching the text as part of our analysis procedure, given that a random or uncontrolled look around is no more likely to work for the larger text than it would for the smaller one represented by the single sentence. For it is unreasonable to suppose that the information we want is scattered about the text in an entirely arbitrary way, given the assumption which we have made throughout, that we are dealing with coherent text: if there are restrictions on the search for significant semantic headings within a sentence, there will probably be restrictions on any search in a more extended text. And the question then is what these restrictions are. In other words, we have to consider the structure of large pieces of text, like paragraphs, as well as that of individual sentences.

Paragraph Structure and Restrictions on Extra-Sentential References

The notion of sentence structure, though it presents many problems, is comparatively well understood. The notion of paragraph or discourse structure, on the other hand, is very ill understood. But some attempt to understand it is nevertheless necessary for any serious work on discourse analysis or interpretation. In my view, one of the most unfortunate consequences of the large amount of work that has been done on sentential syntax is that it is difficult to look at a paragraph as anything other than a string of independent unit sentences, and this inhibits any attempt to think about the overall structure of larger pieces of text. Intuitively there is something implausible about the notion that the structure of a paragraph is simply exhibited by the concatenated structures of its competent sentences: if we look at a paragraph we can perceive relationships between the sentences; and it is this structure, which may be described as the macro-structure of a text as opposed to the micro-structure exhibited in its individual sentences, which we have to examine. I cannot say much about this notion of the macro-structure of larger pieces of text yet; I can only justify my assertion that there is such a thing (if anybody doubts it) by referring to the fact that we can make abstracts, that we can take longer texts like the economic history paragraph and summarize them: thus we might say that this particular text deals with the improvement of road communications in the eighteenth century. This summary has a structure in the sence in which a mere listing of the relevant headings or items would not. If we simply list the headings, ROUTE, COMMUNICATIONS, DEVELOPMENT, TRAFFIC, 18 CENTURY, and so on, we are not summarizing the text: in our genuine summary we are saying that the communications improved, and this statement has a structure in some obvious sense. It seems fairly clear, moreover, that in giving such a summary we cannot but rely on some structure in the original text which must be inferred from the individual sentence; and it is this structure with which we are concerned.

We are concerned with it for two reasons. The first is that we may rely on it in analyzing parts of the text, for example in searching for the information needed to resolve some ambiguity in a particular word; and the second is that this structure must itself be exhibited by any analysis or description, and must be maintained in any transformation of a text, in, say, abstracting or translation. The analysis of individual sentences involves not merely interpreting single words, but exhibiting their relationships, so that, for example in dealing with "white house" we would have to show not merely that "white" means light-colored and not harmless, and that "house" means residence and not assembly, but also that we have a light-colored residence and not a residential lightness. And this requirement holds for larger texts too.

Distinctive Features of Semantic Paragraph Structure

Formal or 'Mechanical' Discourse Organization

Disambiguation Devices

What, then does this notion of discourse structure involve? So far, all we can say is that the existence of summaries or paraphrases which do not consist simply of lists of headings suggests that there is such a thing, and that it is perhaps more natural to describe this thing as semantic than as syntactic structure. One possible approach is to look more closely at the consequences of the need to resolve ambiguity. Suppose, therefore, that we have a sentence with some word in it for which we require more information. If we assume that the author (speaker, writer, or whatever) who is producing a piece of discourse wishes to communicate, i.e., make himself clear, it presumably follows that the ambiguity of any particular item will be resolved as soon as possible. In other words, it seems reasonable, given an ambiguous word, that we should look at the proximate, as opposed to the remote context of our word, which means that we should look at the next sentence if we cannot find what we want in this one. Of course, this picture is oversimplified, and such an approach would probably not work in a given case. For one thing, it ignores the fact that some part of a text will generally have preceded a particular word, so that we may look at a preceding sentence as well as a following one, and secondly, that it masks what may be described as the two-directional feature of disambiguation, namely, that if the presence of word B helps us to select a sense of A, it may also be that the presence of A selects a sense of B. This is clearly exhibited in the repetition model. Therefore we can say that a given word may depend on a subsequent one, in that the latter confirms an interpretation of the former, but equally that the occurrence of the former may assist in selecting the correct sense of the latter. We might construct a very crude model of disambiguation as follows: we tentatively interpret word A; the occurrence of word B confirms this, given a tentative interpretation of B; and now the occurrence of C confirms this interpretation of B, and so on. (The latter confirmation of the interpretation of B is also indirectly one of A.) This description is clearly far too naive: I am merely using this example to illustrate what seems to be involved in the requirement that some information must be supplied to select the correct sense of an ambiguous word: in fact, all that is involved is that for effective communication this information should be provided as soon as possible, or at least sooner rather than later.

Now we can describe this as a 'mechanical' constraint on discourse. If we assume that a piece of discourse is intended to have a theme, or to convey some message, any ambiguity must be dealt with fairly quickly or it will impede comprehension. And this constraint imposes some very minimal structure on a piece of discourse, in that there will be some sentences which are related because one is a clarification of another, where this relationship may be represented by the recurrence of some concept or the occurrence of related ones.

The point which naturally arises now is whether there are any other features of discourse of this kind which have something to do with its structure, and which either may be used for analysis or may have to be taken into account in it.

Linking Devices

Consider, for example, a text like the following:

"General de Gaulle is a big man. There is an extension of influence throughout Europe. The minister has played an important part in the NATO meetings."

One striking feature about this is that there are no connections between the sentences in a text like this:

"General de Gaulle is a big man. The General has extended his political influence throughout Europe. This influence has been especially evident in the part played by the General's minister in the NATO meetings."

Here we have occurrences of the same words like "General" and "influence" in successive sentences. And if we return to our imaginary author, it seems clear that if he wishes to present his message coherently, he must ensure that there is some connection between the parts of it, where these connections or links take the form, say, of repetitions of the same word, so that the reader (or hearer) can follow what is happening. It may be the case that each sentence in a text is related to others because they deal with the same subject: thus "General" occurs frequently in the example just given because the text is about de Gaulle. In this case we would intuitively say that there is an underlying relationship between the different sentences. But these relationships may not be readily perceived unless there is also a surface connection or linking between them which makes this relationship clear. For instance, suppose we have the text:

"General de Gaulle is extending his power. France has played an important part in NATO meetings. Several decisions were taken. Troop quotas are to be reduced. Savings in foreign expenditures will boost the franc."

We may be able to infer what the message of this text is, namely that General de Gaulle was able to cause NATO to reduce troop requirements, thus reducing expenditure in francs outside France. But though we may be able to see what a text like this is about, there is little doubt that we would follow what was happening more readily if the text looked more like this:

"General de Gaulle is extending his power. Thus France has played an important part in NATO meetings. Several decisions have been taken at these meetings. One is that troop quotas are to be reduced. The savings resulting from these reductions will boost the General's franc."

The connections which enable us to follow the discourse are established by a variety of linking devices: the use of pronouns and demonstratives are obvious examples; the repetition of a particular given word or synonym may mark a connection; and syntactic parallels or equivalences may also play a part. These devices may be described as anaphoric, and some work has been done on the problems of identifying anaphoric expressions and discovering their references or antecedents. A variety of such devices exist, all having the same effect, namely that of holding a text together.

Repetitive Devices

Consider now a text which does satisfy the two requirements we have been concerned with so far: provision is made both for resolving ambiguity and for exhibiting the connections between sentences. We again assume that the author's object is to present some genuine or coherent message.

"The boy liked the girl. The girl had a rabbit. The rabbit was white. It liked lettuce. The lettuce grew in the garden. The garden was large. It reached down to the stream. The stream flowed fast. It hurried past the rocks. The rocks were very big. Sometimes they moved. They moved when there were floods. This did not happen very often. But when there were floods they did great damage. They broke bridges and washed away houses. So the houses were very strongly built. They were built of brick on good foundations. The foundations went deep and were made of concrete."

In this case we may find it difficult to understand the text because no one part of it is emphasized. It is not easy to pick out the important point in the face of all the detail, and we are indeed led to doubt whether the text has any specific theme. The further we read, the more difficult it becomes to relate new items to what has gone before as a whole. However, if we rewrite it as follows, we may see more clearly what the real message of the text is:

"The boy liked the girl. The girl had a rabbit, a white rabbit which liked lettuce. The lettuce grew in the garden, which was large. It was so large it reached down to the stream. The stream was a fast-flowing one. It hurried past the rocks. These were big, but they were sometimes moved by the stream when it flooded. This did not happen very often. But when there were floods, they did great damage. The stream sometimes caused floods which broke bridges and washed away houses. Because of the danger from the floods, the houses were strongly built of brick on good foundations, which were deep and were made of concrete."

We can say, when we have read this, that the subject or theme of the text is the stream and its floods, since these words are repeated.

This suggests, therefore, that some degree of repetition is necessary to drive the point of a piece of discourse home; and this suggestion was indeed made by Bally, who said:

"La répétition est une nécessité de la communication et de la pénétration des idées; ce n'est pas chose aisée de se faire comprendre immédiatement en parlant et surtout d'imposer à l'inertie de l'interlocuteur." (See Traité de Stylistique Francaise, Vol. 1, pp. 98-104.)

We thus have a third requirement, namely that some emphasis, which may be achieved by repetition, is needed for efficient communication. None of these requirements can be put very strongly: we saw that to resolve ambiguity we can only ask that the necessary information be supplied sooner rather than later. Equally, it would be unreasonable to say that every sentence must be explicitly linked with the previous one: we can again only say that individual units, sentences, or clauses, or whatever, should be linked to some other nearby item; and similarly, we can only say that emphasis is ensured by some repetition. What we have, therefore, is a set of minimal requirements which must be satisfied if the message of a piece of discourse is to be effectively conveyed. This assumes that there is some such message: and the important point about the mechanical constraints on discourse with which we have been concerned is that it seems that they must be satisfied, whatever the particular theme or message of a piece of discourse is.

Content Organization: Message and Message Structure

In some minimal sense of structure, therefore, these constraints impose a structure on text; and the existence of this structure is clearly relevant to any discourse analysis procedure. It is quite clear, for example, that such a procedure must recognize anaphoric expressions. At the same time, whatever structure there is in a piece of discourse which has a theme or presents some message is not imposed solely by these constraints. I have assumed throughout that the author of a text has some message which he wishes to communicate, and what is more, that this message, since it is a message, has some structure of its own. Without this, a text which simply has the structure imposed by attempts to satisfy these mechanical requirements will not necessarily convey any message, even if the individual sentences are quite sensible. We can construct such a text, for instance, as follows:

"The rose is red. Red is a color. Red is socialist. A socialist believes in giving a government many powers. Power in cars is called horsepower. This is because the tractive power of a car can be compared with that of a horse. Horses are useful animals. There are many different breeds of them. They are mostly brown in color. Some cars are red. I think having cars in many different colors is silly. It's part of the great capitalist conspiracy. Capitalists simply want to make money out of ordinary people; they are not satisfying any genuine needs."

This text substantially satisfies all three requirements, in that provision is made for removing ambiguities, individual sentences are connected with others, and there is some repetition. But the fact that the resulting structure is superficial is shown by the fact that it is not easy to say what the text is about, or whether it has any message which can be summarized. It is not meaningless, because we can assign some semantic interpretation to each of the sentences, so we might say that it has a message in some sense. But it does not convey any message in the strong sense defined by our being able to abstract the whole. In considering discourse structure, that is, I wish to assert that a piece of text larger than a single sentence has a message only if we can produce an intelligible summary of it without reproducing the text itself. What might be meant by the message of a piece of discourse is indeed controversial. For example, take the text:

"The elderly farmer was slowly preparing his field. He drove his two-horse ploughing team backwards and forwards from the wood at the top of his field to the hedge at the bottom, turning the soil in narrow strips."

In one sense, the message of this text is that the elderly farmer was slowly ... etc: it is the contents of, or whatever is conveyed by, the individual sentences; so that we have a description of a message which applies equally to this text or the previous one, and does not indicate any difference between them. But we can intuitively distinguish the two, by saying that the latter is more coherent; we could not produce any very plausible summary of the first text, but we can summarize this one by saying that the farmer was ploughing his field. And, given that text in general appears to be more often like the second example than the first one, my argument here is that we should concern ourselves with the message of a text as something which may be summarized: a summary must depend in some way on the overall structure of the text, and it is the structural features of the text which enable us to give the summary, that interest us from the point of view of discourse analysis.

Now as we have just seen, what I have called the mechanical features of discourse create some structure in a text: the first example above is not wholly lacking in structure; but this structure does not by itself indicate the message of a text. We assumed only that these requirements have to be met by anyone who wishes to communicate some message, and the important point now is that he indeed has to have some message or argument or statement or proposition which he wishes to put across. And obvicusly, since we have defined a message so that it has a structure and does not consist simply of a set of headings, the structure of the message must influence the structure of the text which conveys it in some way. Can we, then, say anything about the structure of messages, or about the ways in which this message structure is exhibited in text? This aspect of discourse structure is clearly far more important than the mechanical one: my point so far has only been to show that at least two aspects of discourse structure may be distinguished, that the structure of a piece of discourse as a whole may be influenced by both of them, and that the existence of this distinction may therefore be relevant when we come to considering analysis procedures.

But what does saying that the message conveyed by a piece of text has a structure amount to? To start with, it must be emphasized that the problems presented by the notions of message and message structure, or of semantic structure in discourse, are the most formidable and intractable of any encountered in the study of language. I cannot attempt, therefore, in what follows, to offer anything like a complete or satisfactory explication of them. All I can do is point to some of the questions which are involved, with the object of showing that the use of a semantic classification seems to be required for any procedure for discourse analysis as it was for sentence analysis.

I introduced the notion of message and message structure very briefly, simply by contrasting the characterization of a piece of discourse by some list of concepts or headings, and by a statement exhibiting relationships between these concepts; and this initial crude notion of message structure should perhaps be developed further before we go into the question of how it may be represented by a text. It is well known to information retrieval specialists, for example, that a simple statement of the topic or subject of a text does not tell us very much (the question of whether it tells us enough is one of the key problems of information retrieval); thus the description of the text about the farmer which simply characterizes the text as being about FARMING is not very informative. Such a description is baseu on the occurrences and recurrences of particular words or conceptual headings: since they recur, it is reasonable to assume that they represent the subject matter of the text. Suppose, however, that we take a single sentence for convenience and imagine it is a complete text: thus we may have "The man bit the dog," We can say that this sentence is about MAN, DOG and BITE (or MEN, DOGS, and BITING, since it is important that the individual words used as headings should not be given too much weight). but this does not tell us what the text is saying. The message of the text is that the man bit the dog, and not that the dog bit the man. When we have a longer text like the one about the farmer, it is not enough to say that it can be described by FARMER and PLOUGHING (or FARM and PLOUGH, or FARM, AGENT and PLOUGH): the message is that the farmer was ploughing in such and such a way, not that the ploughing was farmed. Again, in the earlier text about General de Gaulle, the message can be summarized by saying that France was benefitted by de Gaulle's influence, not that de Gaulle was influenced by France, or any other possible way of relating FRANCE, DE GAULLE (PERSON), INFLUENCE and BENEFIT. The fact that we may intuitively recognize a difference between a description of a text which says that it is about France, de Gaulle, influence, and benefit, and that it is about de Gaulle's influence benefitting France, and that we may say that the difference is that the latter has a structure, does not mean that we have a very clear idea of what sort of a structure this is. The only thing that can be said about it is that it is reasonable to assert that in these cases we have some semantic structure: in a quite crude sense, we can say that the statement that the man bit the dog has a semantic structure, namely that it was the man that bit the dog, that the man did the biting and the dog was bitten, that it was biting that the man did, and moreover, biting of the dog, and so on, even if we can say that the statement has a syntactic structure as well.

The Relation Between Message Structure and Formal Structure

Suppose, then, that we have a very vague idea of what we might mean by the semantic structure of a message. I have already argued that the mechanical and message structure of a text should not be confused, but we should now look at the way in which they may be related. Consider, for example, the following text:

"The President sent for his advisers. He wanted to discuss his proposed educational reforms. These were primarily concerned with widening the curriculum in high schools, and they were the outcome of one of his strongest political convictions, since he regarded education as the key to a better society."

This text satisfies our mechanical requirements; but the fact that it does so appears to be incidental. The fact is that the text is concerned with a particular theme, namely the President's interest in education, and in the course of presenting the message the discourse construction constraints mentioned earlier have been fulfilled. Because the text deals with the President and education we find, for example, the repetition and connections which are also required to maintain the coherence of a text. In other words, the primary influence on the organization is (not unnaturally) what the author wants to say, and this imposes the strongest constraints on the arrangement of the component words and sentences, given that it essentially determines what is said, and how it is said. If we want to go on talking about the President we shall do so, and the fact that one consequence of this is that some of the requirements about repetition and connectivity, say, may be satisfied is accidental. As we have seen, these requirements could not be put very strongly. And we can perhaps reformulate them now by saying that if an author has not satisfied them incidentally in the course of presenting his message or developing his theme, then he must take positive steps to do so. We are all familiar with the use of a concluding sentence in a paragraph which brings out the main point of the preceding stretch of text: this is often not an integral part of the message which has been presented; it is simply a repetition of the chief points and is required to drive them home to the reader. We can, therefore, legitimately describe this as an example of the way in which the mechanical constraints on discourse may have to be fulfilled, although the preceding piece of discourse has presented a genuine message in a coherent way.

Even in a case like this, however, the particular form which this repetition takes will be determined by the message. So even where the mechanical constraints are not already satisfied by the natural presentation of the message, and have to be specially fulfilled, how this is done will be a consequence of the particular message; which means that we can treat the occurrence of some such mechanical device, such as a linking anaphoric expression, as a clue to what the message is. And this is the motiviation behind studies of anaphora, and also of the behavior or distribution of particular words which may be, say, repeated, or have synonyms, though they may not occur anaphorically. Because the various devices we have considered do contribute to the cohesion of a piece of text, and are used to remove ambiguity or to emphasize given points, they cannot but be used in the presentation of the message of the text, and so may be taken as clues to what this message is. At the same time, the two-directional nature of this relationship between device and message will be apparent. If the occurrence, say, of an anaphoric expression is to be taken as a useful clue to the message of a text, as opposed to being a mere connective device, it is also the case that its use in such circumstances is determined by the message. In other words, we may only accept such a clue as being a clue if it points to something that looks like a message.

Discovering the Message of a Text

What we now have to look at is how we discover what the message or semantic structure of a text is. If we take a text like that dealing with the President's interest in education, it is obvious that we must learn something from the syntactic structure of the individual sentences, from syntactic relations between them, from the presence of anaphoric devices, and from the repetition of given words or roots. But can we get everything we want from this information? Suppose that we collect all the pieces of information of this kind which we may get out of the text just mentioned. We might, for example, give all the details of the syntactic structure, according to some suitable model of syntactic description. We might also note, for instance, that various relationships are exhibited by items like "his," "he," "these" and so on and by the recurrence of words like "education." Suppose, moreover, that we have managed by some means or other to select some specific sense for each word. Would we be able to show that we have understood the message of the text by paraphrasing or summarizing it correctly?

My answer is that we probably would, or at least might get some way towards it, but that this is because we are implicitly inferring or relying on, though perhaps only tentatively, the structure of the message, which we have identified by other means, to sort out all the items of information in the text. Essentially semantic considerations appear to enter into syntactic analysis, and as I noted earlier, the effects of, say, connective devices or verbal repetition may only be recognized if we relate them to some possible or hypothetical message. And the question then is what does it mean to say that we have some notion of the message of a text which is not derived from these pieces of information, and in fact on which we rely in interpreting them? In my view, the correct answer to this is that we make use of concepts or semantic markers, and relationships between them.

Using a Semantic Classification to Identify Messages

This point is best brought out by considering disambiguation. We started with the argument that selecting the correct senses of words must depend on semantic headings or markers representing a conceptual classification of word meanings, and further, on permitted combinations of these headings; though the detailed nature of such combinations was not specified then, the main point being that they involved the use of the conceptual classification of words. The important feature of this argument is that the use of such permitted combinations will allow the analysis of sentences containing different words but dealing with essentially the same situation. Thus with the combination EAT FOOD we may be able to handle sentences like "My aunt is chewing candy" or "My aunt is nibbling chocolate" or even "My aunt is munching bacon." If we say that the message of the last of these is that my aunt is munching bacon, then we might describe the combination EAT FOOD as defining a message type or message form: namely, we have individual sentences dealing with various ways of eating various kinds of food, but they nevertheless all share the common characteristic of being about eating food. We can say that the particular different messages may be subsumed under a single message type.

The point is that resolving ambiguity involves the use of a conceptual classification of words, and permitted combinations of semantic headings; and further inspection suggests that the use of semantic headings and combinations of them is involved in obtaining at least some of the information about a text I mentioned earlier. The type of semantic information on which a syntactic description may depend could be of this kind, or at least this is what appears to be involved in, say, features and selection rules as described by Chomsky. Similarly, the correct references of an anaphoric expression, especially of a more sophisticated type, may be determined by such semantic information. For instance, if we have the sentences:

"The members of Smith's team then undertook a long series of elaborate and costly experiments, and published a number of valuable papers about them. These investigations were not, however, followed up by other workers for several decades."

It is at least arguable that identifying the correct referent of "these investigations" as "elaborate and costly experiments" and not "valuable papers" depends on recognizing that "experiments" and "investigations" share a common semantic heading. And, in general, identifying some word as a synonym of another, if both have several uses, depends on the assumption that it is more natural to interpret them as synonyms if they appear in the same text. This is because it is more likely that the same common concept will be relevant in one context than any of the other different headings under which they may be subsumed. Again, the necessary emphasis may be achieved not because individual words are repeated, but because certain ideas such as that of DESTRUCTION associated with "damage," "broke," and "washed away" are in the text about the floods.

It is thus apparent that semantic markers play an important part in governing the use, and permitting the identification, of the different devices we have considered; so that saying that these devices help us to establish the message of a text depends in turn on our identifying the key concepts associated with a piece of text. However, as we have already seen, simply identifying the semantic markers which may characterize a text, however important this is, does not do enough for us, in that it does not give us sufficient information about the text. If a text has a message, it must be reflected by some particular arrangement of these concepts, or in other words by some permitted combination of them, particularly if we regard a permitted combination as structured in some way. The combination EAT FOOD, for example, is assumed to be structured in the minimal sense that the two components are ordered. And it is possible that some more sophisticated structure might be involved in permitted combinations, though I do not want to go further into the details of this question here. Now if we are concerned with such structured permitted combinations, what we are really interested in is message forms; and my argument is precisely that we do discover what the message of a text is by considering the detailed features of it in the light of what we already know about the kinds of things which may be said in a language, namely in the light of message forms based on concepts or general semantic headings, or on the relations between them. Moreover if this hypothesis is correct,* then the importance of investigating semantic classifications becomes much stronger: such a classification is needed not merely for selecting the correct senses of individual words, but also because discovering what the message of a particular text is depends to some extent on a prior knowledge of what types of message there may be, and this in turn depends on some knowledge of the concepts underlying words, that is, on some knowledge of the semantic classes to which they may belong.

But if we say that we use what appears to be the message form underlying a text to sort out its message in detail, we still have to say how we conclude that some particular message form is the relevant one, and we can only infer this from the words. This difficulty is brought out if we get back to the question of extending our search for the information needed to resolve ambiguity outside an individual sentence, and to the problem of what restrictions there may be on this search: resolving ambiguity is not the only purpose for which some knowledge of the message form underlying a text is required, but it is an important one, and it usefully pinpoints the present difficulty. In one sense, the question looks spurious: if we are looking for the components of a permitted combination, we assume that the structure of the text will specify where we look. But the underlying structure of the text is also defined by which permitted combinations occur. Hopefully, we may break out of this circle because though there may be several concepts associated with the words in our text, these may not all combine in an acceptable way. And the tentative suggestion is that this will equally be the case with larger pieces of discourse.

* It is not at all clear how it can be proved to be correct: I can only say that I believe it, and that some such assumption appears to underlie what attempts there have been at semantic analysis.

Message Patterns and Paragraph Syntax

It must, however, be emphasized that new problems appear when we proceed to larger texts, or at least that difficulties which may not be obvious when we confine ourselves to single sentences become so when we look at a paragraph. What, then, are the consequences of extending the notion of permitted combination or message form to larger pieces of text? The first is that it is unreasonable to assume that there will be some single combination or form, of the kind described so far, for a large piece of discourse like a paragraph --though there may well be one for the summary we may make of the paragraph. It is more plausible to imagine that the message structure of a paragraph depends on several or many of these forms, and on the relations between them. But the difficulty which now arises is that this makes for much greater complication in identifying the correct combination of message forms characterizing the whole, and in identifying the particular one which is relevant to a given word in a given sentence. Indeed the crucial problem not only for disambiguation but for discourse analysis in general, is that if we have more places to look for information, this makes the search more of an effort.

It is, unlikely, however, that there are no restrictions on the search. Even within individual sentences, indiscriminate searches for the components of permitted combinations may well pull out false combinations; and it was suggested earlier that the syntax of a sentence may restrict a search in some way. When we now ask whether there are any restrictions on the search for combinations in larger pieces of text, it seems probable that there are, which means that we have to consider the syntax of a paragraph, or what we might call the macro-syntax of discourse as opposed to the micro-syntax of sentences.

This notion of paragraph syntax presents great difficulties: it is not clear that it is exactly like sentential syntax, though it must be related to it in some way; and all I can hope to do is indicate some of its possible characteristics.

To start with, we should say more about what it is we are looking for. Essentially, if we imagine that our individual sentences depend on a single message form, and that a paragraph depends on a number, are there any regularities to be observed in the way in which these message forms may be combined? For it must be recognized that if a piece of discourse is to be coherent, there must be some connection or relation between the component message forms: and the question is whether any patterns may be observed in actual text, or there are any rules about the way in which message forms may be conjoined. For if the combination of message forms in a text did follow any rules whatever, it would clearly make the process of identifying the message forms underlying a text, and their combined structure, much easier. For disambiguation, for example, it would tell us the natural place to look for further information, though we need to know what the message forms of a text are for more reasons than this, as I hope I have made clear. Is there, then, a large syntax in discourse, and if there is, what could it look like?

Topic and Comment

It may be helpful, in considering this question of discourse syntax, to think about it in terms other than those customarily employed for sentential syntax, partly because we are primarily interested, not in the internal structure of sentences, but in the relations between them, and partly because we may well be dealing with a different kind of syntax. One possibility is to look at an individual sentence as putting forward a topic and a comment on it, so that we can describe the relation between them by saying that the comment is legitimate, given the topic: thus if "cow" represents our topic, then "moo" is a legitimate comment on it, while "ratiocinate" is not.* A message form then represents a particular set of similar or related topics ("aunt," "uncle," "sister," etc.) and a set of similar or related comments ("eat," "munch," "nibble," etc.) As a general statement about the surface structure of English or any other sentences this is hopelessly inadequate, though it may be more relevant to deep structures; and of course saying that a comment is legitimate is not saying very much, because we have to define "legitimate" here, or at least list the legitimate kinds of comments on such and such types of topic, and this is simply another way of saying that we have to give permitted combinations. So my justification for using the terms "topic" and "comment," though they are ill-defined, is simply that they are handy terms for discussing intersentential relations, since they are not colored by being used in connection with intrasentential syntax, and they are preferable to, say, "subject" and "predicate." The fact, moreover, that there are many sentences which do not exhibit a simple topic and comment in any very obvious way is unimportant for the moment. As I want to consider paragraph structure--that is, the relation between sentences --there is much to be said, in order to distinguish the wood from the trees, for behaving as if the component sentences in a piece of discourse have a reasonably straightforward internal structure. After all, we do meet sentences like "The car is coming," or "John is happy," or "Mary sings," and we can say that such sentences consist of a topic and a comment. So if we do assume that all sentences are like this, with a view to describing the relations between them, we can then return to the problem of fitting more elaborate sentences into whatever description we can set up.

This terminology brings out an important feature of the notion of paragraph syntax with which we are concerned. This is that it is, so to speak, a semantic kind of syntax. If we assume that individual sentences present topics and comments, any connection of, or relation between, sentences will be either through topics or through comments. The structure of a paragraph must depend on these topics and comments, and topic and comment are,

*The notions of topic and comment are blanket ones, which can and should properly be subdivided: but they can be used as they stand for illustrative purposes.

intuitively at least, semantic notions. The well-formedness of a large piece of discourse does not consist trivially in the fact that it is a string of individually well-formed sentences, the latter well-formedness being characterized in the standard way. In my view, well-formedness in a paragraph is much more a matter of semantic or conceptual coherence: in the last resort, we say that the characteristic property of a paragraph is that it has a recognizable theme, that it deals with some particular idea or set of ideas, and not just a mere hodgepodge of them (we can, after all, make semantic sense of "Colorless green ideas sleep furiously" by embedding it in a paragraph where some particular theme is developed in the light of which it may be interpreted). And we say that a paragraph has a theme, or is semantically coherent, if the individual sentences in it are concerned with the same or related ideas, and to a sufficient extent, or repetitively enough, for it to be quite clear that this is the theme. In this context, it should be noted that our motive for starting a new paragraph is that we have a new theme: we are not governed by the kinds of consideration which determine sentence divisions. Given, therefore, the fact that the topics and comments of the constituent sentences in a paragraph may express the same or related concepts, since they must do this at least if the paragraph is to be accepted as coherent, the question is whether the notion of paragraph structure involves anything stronger than this; does the assertion that a paragraph must have a structure of the kind described mean only that some topics and some comments must be related, or does it imply that topics and comments must be related in specific ways? So the question of whether there are any regularities or patterns in discourse structure resolves itself into a question of whether there are any systematic relations between topic and comments in a piece of discourse, given that there must be some connections.

Illustration of the Topic-Comment Structure of a Text

One approach to the problem just raised is to attempt to exhibit the topic-comment structure of actual text. On a quite informal basis, for example, relying on our own knowledge of a language and ability to understand a text, and making use of all the linking devices as clues, we might attempt to show what the topic-comment structure of some text is, for instance as follows (some allowances have to be made for the fact that the sentences of this text are not so simple that they contain only single topics and comments, though it is taken from a children's encyclopedia in which the text has deliberately been kept fairly elementary):

The items selected as topics and comments are underlined; the different items are numbered, and their occurrences and topic or comment status are noted alongside:

(1) The Pacific is by far the world's largest ocean. 1,T 2,C
(2) Scattered over it are thousands of islands. 1,T 3,C
(3) These islands make stepping stones across the ocean. 3,T 4,C; 2,C

(4)	Boats and planes can stop at them.	5,T	3,C
(5)	The Hawaiian islands make the first stepping stone in traveling across the Pacific from the West Coast.	6(3),T	4,C; 1,C; 7,C
(6)	The step to these islands is a giant one.	4,T; 6,T	8,C
(7)	They are over 2000 miles out in the Pacific from the coast of California.	6,T	9(8),C; 1,C; 7,C

This analysis is obviously highly intuitive and extremely rudimentary: it can be argued, for instance, that relevant items, like "traveling" in sentence 5 have been omitted, which could be associated with "boats and planes" in sentence 4, so that the comment entry for sentence 5 should read 4,C; 5,C; 1,C; 7,C. (But note that such an argument depends on a recognition of the fact that the same concept is conveyed by these different expressions.) And indeed it would be possible to argue about every feature of this analysis. But I still maintain that the mere fact that we can do this kind of thing at all, and that some of the results at least are intuitively acceptable, is grist to my mill; and we may be able to infer some semantic patterns from the study of more examples like this, especially if they have been worked out more carefully. One such pattern, for instance, might be the repetition of the same topic with different comments, as in sentences 1 and 2 here, and also 5 and 7; another might be the use of the comment of one sentence as the topic of the next, as in sentences 2 and 3, and perhaps 5 and 6. In analyzing this text I relied chiefly on the more obvious ways of relating one item to another, namely, pronouns, demonstratives, and repetitions of the same word but one case where there is a link in terms of general semantic headings occurs in sentences 6 and 7 for "giant" and "2000 miles." This emphasizes the fact that we should always look at the devices as evidence of conceptual relationships; it is often the case that the choice of what particular device we use is arbitrary, so long as the conceptual link is preserved: thus we could have "The Pacific" instead of "it" in sentence 2: but there may be a conceptual relation between two sentences, without there being any very obvious connecting device, as in sentences 6 and 7.

It is nevertheless clear that it may be difficult to pinpoint any well-marked patterns in these analyses; and the vital question from the point of view of discourse analysis is whether, interpreting structure as patterns of topics and comments, we can see any particular patterns, and whether we can say, as a result, that discourse construction is governed by the rules defining these patterns. It has been persuasively argued that we cannot hope to do this, simply because the particular structure of a given paragraph is determined primarily by what its author wants to say, and not by how he ought to say it. In other words, if we have two successive sentences with the same topic and different comments, this is because the author wanted to go on talking about the topic in question, and not because he was following a discourse construction rule which says that he must repeat his topic. In this case, the only constraints on the discourse structure are those imposed by the need to maintain a certain degree of coherence by sticking to a theme (or, to put the point less strongly, are the consequence of the fact that we tend to stick to a theme); and as we have seen, such structure as may be imposed by this conceptual selection may not be very easy to use, or of much value, when it comes to identifying the specific message forms underlying the text.

Argument Development

At this point, however, we can introduce an idea which has not been mentioned explicitly so far, namely that of argument development. We have behaved so far as if a text is semantically coherent if it deals with a particular theme or set of concepts, without there being any question of one sentence following another, or of a subject being developed; we have talked about a paragraph as if, metaphorically speaking, we could see (or hear) it all at once; but we do not normally pick up a piece of discourse as a whole. What is more important, we do not write a paragraph as a whole. So the notion of coherence in a paragraph may be interpreted more strongly, to imply not merely that the constituent sentences should share common concepts, or should refer to related concepts, but that the way in which they share them should be associated with the order of the sentences. We do talk about the thread of an argument; we do say that one sentence follows semantically from another: and one would therefore expect that the organization of the message forms for a piece of text, in terms of conceptual repetitions and relations, would exhibit the development of a theme. A piece of discourse is dynamic in this sense, and this must influence what we mean by discourse structure. What we can now ask, therefore, is whether the fact that there is some progress from the beginning to end of a piece of discourse, the fact that there is some development of an argument, does put constraints on discourse structure over and above those imposed by the general need to maintain coherence. To put the question in its simplest form, does what we have said already influence what we say next in any significant way? Because if it does, then we may be able to draw some conclusions about patterns in discourse.

Can we, then, say anything about argument structure in discourse in general? Unfortunately, remarkably little attention has been devoted to the nature of argument in ordinary language discourse. In general, such attempts as there have been to analyze the notion of argument have been concerned with the nature of a valid or true argument, and with defining types of valid argument, or modes of logical inference. The object, given individual statements in a special topic comment form, is to define rules for linking these statements by their topics or comments in such a way that a final statement is obtained which is guaranteed to follow from the initial ones in the strong sense that if these are true, then the final statement is true. For example, one way of looking at the syllogism is to lay it out:

A is B
B is C
A is C,

so that the relations between the tcpics and comments of the statements in question are plainly shown.

But the historical development of this line of thought shows what is wrong with it from our point of view. It has led to the analysis of the nature of proof in formal systems, and the obvious difference between a mathematical argument with its terminating Q.E.D. and the kind of argument which we meet even in what we would describe as closely reasoned discourse suggests that the logician's approach will not help us very much. The difference between logical demonstration on the one hand and the presentation of some subject in ordinary text is that we do not usually ask, at the end of an ordinary paragraph, whether the last sentence follows conclusively from the first. One of the important differences indeed between logical argument and ordinary discourse is that in the latter there is usually no question of establishing a conclusion in the minimum number of steps, as opposed to presenting a subject from a number of different points of view. The basic difficulty is that logicians have concentrated on the kind of discourse which presents an argument in a strong sense of the word, while we have to consider a variety of different kinds of discourse, which present an argument only in a much weaker sense. It is possible that "argument" is too strong a word for our purpose, but it is appropriate because it suggests the idea of development, and this is what is important: we say that a theme is developed even in such apparently unlikely kinds of text as descriptions. And our problem, therefore, is that we have to allow patterns in, say, a report of a director's meeting, a denunciation of a devious politician, or an analysis of the economic problems of underdeveloped countries, and to show what these are. It may indeed be that no such patterns can be found. The only argument that there may be in favor of the hypothesis is that even in such an apparently nonargumentative piece of prose as a description of the Grand Canyon, we can say that successive ideas are presented, and can see that one sentence follows semantically from another. And from this it follows that there must be at least some constraints on what comes after, which are imposed by what has gone before; so that at any particular point in a piece of discourse, we may be able to eliminate some possible message forms in favor of others: and if we can do this, we gain in attempting to identify the structure of a text.

As I said, very little attempt has been made to go into the question of paragraph syntax, so that much of the foregoing is necessarily highly speculative. Nevertheless, two points may be mentioned. One is that the development of an argument may be marked, and indeed may be determined, by what may be called argument devices, in contrast to the mere connecting devices discussed earlier: such expressions as "then," "therefore," "if," "because" and so on have a well-recognized function in this connection, and a study of their behaviour should contribute to an understanding of discourse structure. At the same time, it must be emphasized that discourse structure does not depend only on these devices: it will be evident from what I have said already that conceptual relationships between topics and comments are essential to discourse structure; and my other point is indeed connected with this. This is that these conceptual relations may be highly complicated and various: thus, for example if we take the sentences:

"I spent at least three-quarters of an hour waiting for my sister-in-law who was buying a wholly unnecessary dress. Economy was never one of her strong points, and consideration for others isn't either."

The relations between "wholly unnecessary" and "economy," and between "I spent at least .... waiting," and "consideration for others" are not particularly simple ones. What we have here, in fact, are the ramifications of the notion of permitted combination which was originally introduced for single sentences. And in this context, I should say that in my view we must face the use of these 'loose' connections squarely: there is no future in saying that the structure or ordinary discourse is like a logical proof because if you work hard enough you can spell out all the premises on which such loose connections are based and so make the connection a tight one.

Example of a Simple Model of Discourse Structure

Since no one else has made any very concrete suggestions about discourse structure as we have now interpreted it, and since I cannot say anything about it yet, I shall attempt to make what I have been saying more clear by describing a very crude and indubitably defective model of discourse structure, simply as an illustration. This model is derived from some work which has been done at the Cambridge Language Research Unit. It must be emphasized that the discussion is intended only as an example, and that I am deliberately simplifying the model, and omitting some features of it which are not immediately relevant in the present context. As a model of real discourse analysis, therefore, this would be woefully inadequate.

The Language Research Unit has developed a system of semantic markers or headings consisting of 50 elements, MAN, THINK, DO, CAUSE, CHANGE, WORLD, UP, POINT, WHERE, WHEN, HOW, KIND, STUFF, BANG, etc., from which permitted combinations can be formed, which have a structure given by one or other two relations, ':', and '/', very roughly interpretable as indicating the association of two elements A and B, and the action of one element on another. We may thus set up a list of message forms, containing different elements combined in different ways, as, for example:

MAN/DO
THING/CHANGE
STUFF:KIND
STUFF:HOW
DO:UP etc.

We might imagine our topic-comment sentences, as it might be

The boy ran.
The plant grew.
Cloth is red.,

which would be characterized by such message forms. If a word is ambiguous, it may be described by different elements, and will hopefully have its correct sense identified by the fact that it will fit, together with its sentential companion, in only one form. We might indeed get alternative forms for individual sentences, but we would hope to eliminate these in the course of establishing the structure of the text in which such a sentence appeared. We now specify semantic patterns which deal with, say, pairs of triples of sentences, by relating their topics or comments. Thus, given one sentence with T1 and C1, we will infer that the next sentence will share T1 or C1 with it, where sharing a T or a C means that the actual words in the sentences may be described by the same semantic element. For example if we have the two sentences

The boy ran.
The man walked.,

we have two topics which share the same element, MAN, and two comments which also share the same element, DO. We may further say that if two sentences share the same element this may characterize the comment in one and the topic in the other, so that what is the topic in one case may be the comment in the other, and vice versa. Within this general framework we may have a variety of specific patterns: thus from the last specification we may have

1,T	2,C
2,T	3,C.

Other possible patterns are:

1,T	2,C
1,T	3,C
1,T	4,C;

1,T	2,C
3,T	2,C
4,T	2,C;

1,T	2,C
1,T	3,C
3,T	4,C
3,T	5,C

1,T	2,C
2,T	3,C
2,T	4,C
4,T	5,T
5,T	6,C
5,T	7,C

Suppose then, for example, that we have an initial sentence in which T1 is characterized by MAN, and C1 by DO; we may then see if any message forms containing MAN or DO could characterize the next sentence; if they could, these are prima facie candidates as the correct message forms for this sentence. This would follow, for example, if we had the two sentences

The man is running.	MAN/DO
The running is fast.	DO:HOW.

Now taking the second sentence we consider the possibilities to which it leads: namely that the next sentence may be characterized by message forms involving DO or HOW (and possibly, more elaborately, MAN as well). We then see if any of the possible interpretations of the words in it would fit these as might be the case if the following sentences were "The running is down," or "It is faster than wind." Supposing, moreover, that we in fact have alternative possible message forms for a particular sentence, because the words in it are ambiguous: in this case one may be eliminated because its characteristic elements may not be associated with any word in the following sentences.

One interesting way of using such models is to see whether discourse can be constructed with them, as opposed to analyzed. For this purpose, the statements about accepted patterns would be reformulated as rules: thus we would say that the topic or comment of one sentence must be used as the topic or comment in the next. The justification for this kind of exercise is that if sensible discourse can be constructed by such means, it may be that sensible discourse can be analyzed by them. And experiments, though of a very rudimentary kind, have been carried out on this basis at the Cambridge Language Research Unit: One interesting result is that following such rules with a vocabulary of words suitably characterized tends to restrict the choice of words in each sentence increasingly until, as it were, everything has been said about such and such a theme, and no more can be produced without repetition.

This model of discourse structure, as I have presented it, is so crude that it can scarcely be called a model at all. But it is useful because it is sufficient to illustrate what a message form might be like, what its relation to the actual message of a piece of discourse might be, what the syntax of a paragraph might look like, and what the sementic structure of a paragraph, namely the organization of its particular message forms within its syntactic structure, might be like. In particular, though the model as a whole is inadequate, it is possible that its components may be more realistic than they appear to be at first sight: and since it is very difficult to discuss the problem of semantic classification, which is my primary objective, in the void, without some reference to its background or purpose, I shall assume that discourse structure, at least in broad outline, has the properties I have discussed, and that any discourse analysis will rely on them more or less in the way that I have indicated. In this context, it should be pointed out that though my model appears to be inadequate because it is being applied to unnaturally simple text, this is not such a serious defect as it might appear to be. At first sight, it appears that any attempt to analyze real text will have to be more complicated, given that we are liable to meet sentences like this:

"When a given territory changes hands, the spoken language of the former inhabitants may give way to that of the newcomers, but the place-names normally remain as a perennial monument to the people who first lived there, though they may change to the point where they are practically unrecognizable, like the Celtic or pre-Celtic Eboracum that ultimately became York."

But it may nevertheless be the case that we may be able to discover what the underlying semantic pattern of a text like this is, so that we can describe it in a way which is much more like the model than might apparently be possible, if some sort of fragmentation procedure is applied to it. One suggestion for spoken discourse is that it should be divided into intonational or breath units, which are generally smaller than sentences. Another is that sentences should be reduced to their simple kernel subunits. For instance we might have a set of kernels, with the links between them suitably marked, for the middle part of the sentence given more or less as follows:

place-names remain
remain is normally
remain is as monument
monument is perennial
monument is to people.

This is after all what is suggested by Chomsky when he argues that the semantic interpretation of a sentence is applied to its deep structure and not its surface structure. Thus for example, the sentence "He is very agitated since your PSQ form has not been completed" would be syntactically analyzed in such a way that it broke down (very roughly) into two components representing "He is very agitated" and "Your PSQ form has not been completed." and if we can break down complicated sentences in this way, it may be more easy to see the semantic relations and structure in the text in question.

Conclusion

It will be evident that many questions about semantic discourse analysis have not been answered in this discussion, and that finding suitable answers is not easy. My object in the foregoing was primarily to raise some of these questions, and to try to sort out some of the different features of discourse with which a semantic analysis procedure must be concerned. Moreover, though this presents many problems, the conclusion that a semantic classification plays a vital part in the process of discourse analysis as a whole, and not merely in disambiguation, seems inescapable; and this provides both a justification for investigating semantic classifications and a basis for evaluating them.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* System Development Corporation Santa Monica, California	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Notes on Semantic Discourse Structure

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(Last name, first name, initial)*

Sparck Jones, Karen

6. REPORT DATE 3 March 1967	7a. TOTAL NO. OF PAGES 30	7b. NO. OF REFS
8a. CONTRACT OR GRANT NO. Independent Research b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) SP-2714 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

The semantic problems of natural language discourse analysis are more serious than the syntactic ones, and much less work has been done on them. It is widely held that a semantic classification of words by their meanings is required for selecting the correct senses of individual words in text, but the implications of this hypothesis have not been fully worked out. It is moreover evident that semantic discourse analysis does not end with disambiguation: understanding a piece of discourse depends on identifying its message, and it is arguable that for this we have to know what the semantic structure of the text as a whole is, where this may not be very directly related to the syntactic structure of its component sentences. It is suggested that a semantic or conceptual classification plays a part here, since understanding the message of a text involves some knowledge of the way concepts may be or are usually combined. In this paper, different aspects of this notion of semantic discourse structure are explored, and their connection with a semantic classification is exhibited.

Unclassified
Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Semantic Discourse Semantic Classification Discourse Analysis						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER. If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY. Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (*TS*), (*S*), (*C*), or (*U*)

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14 KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified

Security Classification